SRC TR 89-10

Optimal Scheduling with Strict Deadlines

by

P. P. Bhattacharya

and

A. Ephremides

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
1989

2. REPORT TYPE

3. DATES COVERED
00-00-1989 to 00-00-1989

4. TITLE AND SUBTITLE
Optimal Scheduling With Strict Deadlines

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Maryland,Systems Research Center,College Park,MD,20742

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE
unclassified	**unclassified**	**unclassified**

17. LIMITATION OF ABSTRACT

18. NUMBER OF PAGES
25

19a. NAME OF RESPONSIBLE PERSON

OPTIMAL SCHEDULING WITH STRICT DEADLINES *

Partha P. Bhattacharya, Anthony Ephremides

Electrical Engineering Department and Systems Research Center

University of Maryland, College Park, MD 20742

ABSTRACT

We consider the problem of dynamic scheduling of customers (messages) in time-critical environments. First, we consider a single station (communication node) and assume that each customer (message) must begin service (transmission) by an individually varying "extinction" time or, else, it is lost. We are interested in minimizing, in the sense of stochastic order, the number of messages lost over any time interval. We prove a variety of results that establish the optimality of the STE (Shortest-Time-to Extinction) policy under rather general conditions. Similar results are also shown when messages have constraints on their complete transmission times. If the scheduler is allowed to take decisions based only on the distribution of the deadlines (rather than their exact values), similar but somewhat stronger results are proven. Finally, we consider a network of M stations in tandem under the hypothesis that a message is never lost and is scheduled irrespective of whether its extinction time (also called due date in this case) has expired or not. Again,under fairly general assumptions on the arrivals, deadlines and services, we show that the EDD (Earliest Due Date) policy minimizes a form of average tardiness incurred over a finite operating horizon among all nonidling, nonpremptive policies. We formulate these problems in the context of stochastic dominance, and use simple interchange arguments to establish all our results.

* This work was supported by ONR Grants N00014-85G-0207 and N00014-88K-0712 and an NSF Grant CDR-85-00108.

I. INTRODUCTION

We consider the problem of scheduling the transmission of messages over a single communication link when each message has constraints on its waiting time or complete transmission time. This problem arises in applications that involve time-critical message contents. We wish to model situations in which the penalty incurred when deadlines are not met implies either the complete loss of the message or another form of tardiness cost. We are interested in characterizing the scheduling strategy which minimizes a cost function that reflects the nature of the penalty incurred. The models and the results of the paper apply equally well to numerous other applications that involve service stations, queues and deadlines. Thus a much more general terminology could be used. We choose to stick with the message transmission application in order to focus attention to the important problem of real time communication.

First we consider the case in which the messages have constraints on their waiting times. Each message upon its arrival at time t_i "announces" a deadline d_i, so that if by time $e_i = t_i + d_i$ (called its "extinction time"), transmission does not *commence*, the message is considered lost and never scheduled for service. The objective is to find a scheduling policy which minimizes the average number of lost messages over any time interval. We show that under nonexplosive, but otherwise arbitrary, arrival and arbitrary deadline processes, and for exponential service (i.e. transmission) times that are independent of each other and of the arrival and deadline processes, the policy of scheduling the eligible customer with the Shortest Time to Extinction (denoted by STE) is optimal among all nonpreemptive and nonidling policies. In fact, we show the optimality in the sense of stochastic order. When considered over the broader class of only nonpremptive policies, the optimal policy, if it exists, can be found in the class of STEI policies, namely those that are allowed to idle, but schedule according to the STE rule when they don't idle. As a special case, in the situations in which deadlines are deterministic and identical for all messages, the pure STE policy is optimal within the class of nonpreemptive policies.

Next, we consider the case in which messages have constraints on their complete transmission times rather than on their waiting times prior to service commencement. A message is now considered lost if it does not *complete* transmission by its extinction time. We assume that the transmission of a message is aborted if its deadline expires while it is in the process of being transmitted. Scheduling results that are similar to those of the previous case, are also obtained for this case. Preliminary versions of these results were first presented in [15].

An interesting twist to the above problem is obtained if we assume that due to implementational difficulties, the scheduler doesn't have information about the exact deadlines of the messages. In addition to the knowledge of the past evolution of the system, only the knowledge of the distribution of the deadlines is available for decision making. However the extinction times become known as a message is read and the scheduler can act as before. We assume that the deadlines form a sequence of i.i.d RV's (that are also independent of the arrivals), and the common distribution has non-decreasing failure rate. Service times are also assumed to form an i.i.d sequence of RV's that are independent of the arrivals and deadlines. When deadlines are to the beginning of service, the policy of scheduling the message which has waited the most, is shown to be optimal within the class of non-preemptive policies. Exponentiality of service times is needed further when the deadlines are to the completion of service.

These results do not seem to be easily extendible beyond the case of a single link. However, under a slightly different set of assumptions and operating conditions, some results can be obtained for a tandem network of links. Specifically, we may assume that no messages are discarded or lost and instead, all messages are scheduled, regardless of whether their deadlines have expired or not. A penalty is, however, incurred when a deadline is missed. The penalty function is of the form $h(c_i - e_i)$ where c_i is the time at which the message arriving at t_i with deadline d_i (and extinction time e_i, which in this case is also called due time) completes transmission in the network, and h is a real valued continuous convex function with $h(x) = 0 \, \forall x \leq 0$. For this system, we consider a finite operating time horizon and wish to obtain a scheduling policy that minimizes the total penalty function (usually called tardiness). Under nonexplosive, but otherwise arbitrary, arrival and arbitrary deadline processes, and independent identically distributed service times that are also independent of the arrival and deadline processes, we show that the policy which schedules, at each node, the message with the Earliest Due Time is optimal among all non-preemptive and non-idling scheduling policies.

These problems fall in the category of single server queueing systems with impatient customers. Such systems have received moderate attention in the queueing literature. Most of the work seems to have focussed on the evaluation of various criteria of performance when service is assigned according to a First-come-first-serve (FCFS) strategy. For example in [1,2], the authors are able to compute various performance indices such as steady state probability of rejection of a customer, average number of customers served before the loss of the first job etc. for a FCFS single server queue. In [3], delay analysis is given for a single server queue with an interesting delay-dependent service discipline; this models telephone call processing systems.

While intuition suggests that assigning service according to a FCFS discipline is probably not the best thing to do, attempts to discover better scheduling stragies have been rare. In [4,5], the problem of optimally scheduling the service of impatient customers was posed and weaker versions of some of our results were proved with a different technique and under different assumptions relative to ours. In [6], the problem of optimal control of arrivals to a FCFS single-server queue was considered with the objective of minimizing the discounted reward associated with the successful departure of customers. An optimal control problem with interacting service stations and impatient tasks was studied in [16]. The authors assumed that the extinction times are not known to the controller and the residual time (until extinction) is exponential. The service assignment strategy minimizing a (discounted) weighted average delay was characterized. Several researchers [7,8,9] in the operations research area have considered scheduling impatient customers under various tardiness criteria. In contrast to our formulation, these works assume that the arrival process is shut off after the system has started the scheduling operations and prove results only for a single node.

The nature of the problem we consider makes it difficult to formulate a dynamic programming recursion. Our approach, instead, involves interchange arguments together with the ideas of coupling and stochastic dominance and is based on sample-path wise comparisons of the costs under different policies. Somewhat similar arguments were given in [10,11] for queue-control problems of a different nature. However, since results regarding the existence of optimal stationary policies are not readily available for our problem, some additional care is needed.

The paper is organized as follows. In section II, we introduce the notation and consider the first situation in which messages have constraints on their waiting times. The case in which constraints are placed on their total transmission times is considered in section III. In section IV, we analyze optimality under the reduced information structure. Finally in section V, we present the case of the tandem network with tardiness cost.

II. CONSTRAINTS ON WAITING TIMES

We consider a single server queueing system with unlimited buffer space size that represents a single link of a communication system. Let t_i be the arrival instant of the i^{th} message whose deadline is d_i. We define by $e_i = t_i + d_i$ the extinction time of that message; that is, if by time e_i, the transmission does not commence, the message is considered lost and never scheduled for service. At any instant t, a message with extinction time e_i is termed eligible for transmission if $e_i - t > 0$. Let $\{T_i = t_i - t_{i-1}\}_{i=1}^{\infty}$ (with $t_0 = 0$) be the sequence of interarrival times and S_i be the duration of the i^{th} service time in the system. By service time, we mean the transmission time of a message which may include processing and propagation times as well.

We make the following assumption throught the paper :

(A1) $\{S_i\}_{i=1}^{\infty}$ is a sequence of independent identically distributed RV's which are independent of $\{T_i\}_{i=1}^{\infty}$ and $\{d_i\}_{i=1}^{\infty}$. Also, the arrival process is nonexplosive, that is, $\lim_{i\uparrow\infty} t_i = \infty$ with probability 1.

Let $E(t)$ denote the (increasing) ordered set of extinction times of eligible messages at time t. Let $H_a(t)$ denote the set of all arrival instants by time t and $H_d(t)$ the set of corresponding deadlines. Also, let $C_s(t)$ be the expended portion of the service in progress at time t and $i(t)$ be the condition of the server at time t (1 if busy, 0 if idle). Then, under assumption A1, $z(t) = (\,E(t), H_a(t), H_d(t), C_s(t), i(t)\,)$ is a useful description of the system. Let Z denote the allowable range of values of $z(t)$.

The control action is to decide, at appropriate decision instants, whether to transmit and, if so, which message out of the currently eligible pool of messages. First, we restrict attention to nonanticipative policies throughout this paper. This means that the control action has to be based only on the past evolution of the system, specifically, the knowledge of the service times of the messages waiting in the queue is not available. Let Γ_0 be the class of nonpreemptive and nonidling policies and Γ_1 be the class of nonpreemptive policies, while Γ is the global class of all possible scheduling policies.

For every policy in Γ_0, the decision instants are the instants of service completion (provided that $E(t)$ at these instants is nonempty) or of arrivals to an empty queue. Denote by STE, the policy in Γ_0 which at every decision instant schedules the eligible message with the shortest time to extinction. Let STEI denote the class of policies in Γ_1 which are allowed to possibly idle when messages are waiting in the queue, but which schedule according to the STE mechanism when they choose not to idle.

Following the standard notation, we say that a RV X is stochastically smaller than a RV Y, and write $X \leq_{st} Y$ if $P(X > z) \leq P(Y > z)\, \forall z \in I\!R$. Order relations for stochastic processes can be considered as an extension of the definitions for vector valued RV's. Let $X = \{X(t), t \in \Lambda\}$ and $Y = \{Y(t), t \in \Lambda\}$ be two processes, where $\Lambda \subset I\!R$. Let $D \overset{\text{def}}{=} D_R[0,\infty)$, the space of right continuous functions from $I\!R_+$ to $I\!R$ with left limits at all $t \in [0,\infty)$ be the space of their sample paths. We say that the process X is stochastically smaller than the process Y, and write $X \leq_{st} Y$

if $P\{f(X) > z\} \leq P\{f(Y) > z\} \forall z \in I\!\!R$, where $f: D \mapsto I\!\!R$ is measurable and $f(x) \leq f(y)$ whenever $x, y \in D$ and $x(t) \leq y(t)$ $\forall t \in \Lambda$. The following equivalence [12,13] often provides an easy way to prove stochastic order relations without explicit computation of distributions:

1.$X \leq_{st} Y$

2. $P(g[X(t_1), \cdots, X(t_n)] > z) \leq P(g[Y(t_1), \cdots, Y(t_n)]) > z)$ for all $(t_1, \cdots, t_n)$, all z, all n, and for all $g: I\!\!R^n \mapsto I\!\!R$, measurable and such that $x_j \leq y_j, 1 \leq j \leq n$ implies $g(x_1, \cdots, x_n) \leq g(y_1, \cdots, y_n)$.

3. There exists two stochastic processes $\bar{X} = \{\bar{X}(t), t \in \Lambda\}$ and $\bar{Y} = \{\bar{Y}(t), t \in \Lambda\}$ on a common probability space such that $\mathcal{L}(X) = \mathcal{L}(\bar{X})$, $\mathcal{L}(Y) = \mathcal{L}(\bar{Y})$ and $\bar{X}(t) \leq \bar{Y}(t)$ $\forall t \in \Lambda$ *a.s.* Here $\mathcal{L}(.)$ denotes the law of a process on the space of its sample paths.

Returning to our problem, let $L^\pi(z)$ denote the process $\{L_t^\pi(z), t \geq 0\}$, where $L_t^\pi(z)$ is the number of messages lost by time t when starting from state z at time 0 and applying the scheduling policy π.

We first consider optimality within Γ_0.

Theorem II.1. *Consider a single server queue under assumption A1. Assume further that the common distribution of the service times is exponential. Then, the STE policy minimizes in the sense of stochastic order, the number of messages lost by any time among all policies in the class* Γ_0*, that is,*

$$L^{ste}(z) \leq_{st} L^\pi(z) \qquad \forall \pi \in \Gamma_0, \forall z \in Z.$$

Under the assumption of Poisson arrivals rather than exponential services, STE policy was shown in [5] to maximize the long run expected fraction of successful messages among all stationary policies. Using the fact that the initiation of busy periods are regeneration points, the conclusion followed from a comparison of the quantities for the first busy period. Here, we prove a stronger result using different techniques. We first need the following result:

Lemma II.1. *Consider a single server queue as in Theorem II.1. Let an arbitrary policy* $\pi \in \Gamma_0$ *act on the system in* $[t_0, \infty)$*, where* t_0 *is an arbitrary decision instant. Then there exists a policy* $\hat{\pi} \in \Gamma_0$ *that schedules the customer with the shortest time to extinction at time* t_0 *(and is appropriately defined in* $[t_0, \infty)$*) and satisfies*

$$L^{\hat{\pi}}(z) \leq_{st} L^\pi(z) \qquad \forall z \in Z.$$

Proof. Assume that π does not schedule the customer with the shortest time to extinction at time t_0. If it does, the result follows trivially by letting $\tilde{\pi}$ to be the same as π. We drop z from $L_t^{\pi}(z)$ and $L_t^{\tilde{\pi}}(z)$ for notational convenience.

The idea of the proof is to define $\tilde{\pi}$ appropriately in $[t_0, \infty)$ and to construct two coupled processes $(L_t^{\tilde{\pi}}, \bar{L}_t^{\pi})$ on the same (given) probability space so that $\bar{L}_t^{\pi}$ and L_t^{π} have the same distribution and $L_t^{\tilde{\pi}} \leq \bar{L}_t^{\pi}$ $a.s \, \forall t \geq t_0$.

Suppose $E(t_0) = \{e_1, \cdots, e_n\}$ with $n \geq 2$. We agree to denote by e_i either the extinction time or the message with that extinction time. Let π schedule e_k $(2 \leq k \leq n)$ at time t_0. We will construct $\tilde{\pi} \in \Gamma_0$ which schedules e_{k-1} at time t_0 (and is appropriately defined in $[t_0, \infty)$)and satisfies the assertion of the lemma. The required policy $\tilde{\pi}$ can then be generated by induction on k.

Consider the system evolving under policies $\tilde{\pi}$ and π. Couple the realizations by giving them the same arrival and deadline processes in $[t_0, \infty)$. Let σ be the completion instant of the service which begins at time t_0 under $\tilde{\pi}$. Take the service under π to end at σ as well. This is permissible since the service times are independent and identically distributed. Three cases exhaust the possibilities:

Case 1 : $\sigma \geq e_k$

In this case, under both policies, all messages eligible at time t_0, except e_k for π and e_{k-1} for $\tilde{\pi}$ that have extinction times less than or equal to σ are lost and so are all arrivals in $(t_0, \sigma]$ whose extinction times are similarly less than or equal to σ. The states under π and $\tilde{\pi}$ are therefore matched at time σ. In $[\sigma, \infty)$, define $\tilde{\pi}$ to be identical to π; this is possible since we may take corresponding service times to be equal under π and $\tilde{\pi}$. Thus,

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [t_0, e_{k-1}) \cup [e_k, \infty); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, e_k). \end{cases}$$

Case 2 : $\sigma < e_{k-1}$

First it is clear that $\bar{L}_t^{\pi} = L_t^{\tilde{\pi}}$ $\forall t \in [t_0, \sigma)$, and at time σ, the sets of extinction times under π and $\tilde{\pi}$ differ only in that e_{k-1} is included in that set under π as compared to e_k under $\tilde{\pi}$. Let $\tilde{\pi}$ follow π for $t \geq \sigma$ except that it schedules e_k when (and if) π schedules e_{k-1}. Thus $\tilde{\pi}$ is well defined in $[\sigma, \tau)$, where τ is the end of the current busy period under π.

Suppose π eventually schedules e_{k-1}, that is, suppose that e_{k-1} meets its deadline under π. Since $e_{k-1} < e_k$, e_k will also meet its deadline under $\tilde{\pi}$; and the states under π and $\tilde{\pi}$ are matched at time τ. Letting $\tilde{\pi}$ follow π in $[\tau, \infty)$, one thus obtains, $\bar{L}_t^{\pi} = L_t^{\tilde{\pi}} \, \forall t \in [\sigma, \infty)$.

Suppose now that π does not manage to eventually schedule e_{k-1} before its expiration. Since π is nonidling, we have necessarily that $\tau \geq e_{k-1}$. If $\tau \geq e_k$, then e_k is lost under $\tilde{\pi}$. Letting $\tilde{\pi}$

follow π in $[\tau,\infty)$, one obtains,

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [\sigma, e_{k-1}) \cup [e_k, \infty); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, e_k). \end{cases}$$

If however, $\tau < e_k$, then at time τ, the queue is empty under π and e_k is the only eligible message under $\tilde{\pi}$. Let $\tilde{\pi}$ begin serving e_k at time τ. We must now consider the following two sub-cases:
(a) Suppose there are no arrivals while e_k is in service under $\tilde{\pi}$. The states are then matched from the instant message e_k finishes service under $\tilde{\pi}$; thus,

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [\sigma, e_{k-1}); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, \infty). \end{cases}$$

(b) Suppose there is at least one arrival while e_k is in service under $\tilde{\pi}$. Let the arrival which begins service first under π have extinction time e_b. Take the service time of e_b under π to be equal to the residual service time of e_k under $\tilde{\pi}$. This is possible because of assumption A1 and the memoryless property of the exponentially distributed service times. In this way, we ensure that e_b and e_k will finish service at the same time instant σ_1. Suppose that $\sigma_1 \geq e_b$. Then e_b is lost under $\tilde{\pi}$ and the states under the two policies are matched at time σ_1. One thus concludes that,

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [\sigma, e_{k-1}) \cup [e_b, \infty); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, e_b). \end{cases}$$

Suppose now that $\sigma_1 < e_b$. Let $\tilde{\pi}$ follow π for $t \geq \sigma_1$. Take the corresponding service times under π and $\tilde{\pi}$ to be equal. Thus $\tilde{\pi}$ is well defined in $[\sigma_1, \tau_1)$ where τ_1 is the end of this busy period under π. If $\tau_1 \geq e_b$, we let $\tilde{\pi}$ follow π from time τ_1 onwards, and thus have, the same relations between $\bar{L}_t^{\pi}$ and $L_t^{\tilde{\pi}}$ as in the situation just described. If instead $\tau_1 < e_b$, let $\tilde{\pi}$ begin serving e_b at time τ_1. We are now back to a situation we have described previously. One easily repeats the arguments to obtain,

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}}, & \text{if } t \in [\sigma, e_{k-1}); \\ L_t^{\tilde{\pi}} + 1, & \text{if } t \in [e_{k-1}, \tau_1); \end{cases}$$

and $\bar{L}_t^{\pi} \geq L_t^{\tilde{\pi}} \;\; \forall\, t \in [\tau_1, \infty)$.

Case 3 : $e_{k-1} \leq \sigma < e_k$.

It is clear that

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}}, & \text{if } t \in [t_0, e_{k-1}); \\ L_t^{\tilde{\pi}} + 1, & \text{if } t \in [e_{k-1}, \sigma), \end{cases}$$

and at time σ, under $\tilde{\pi}$, message e_k is eligible for service in addition to all messages that are eligible under π. Consequently, we can proceed as in case 2. We thus conclude that

$$\bar{L}_t^{\pi} \geq L_t^{\tilde{\pi}} \qquad \forall t \in [\sigma, \infty).$$

The observation that the processes $\bar{L}^{\pi}$and L^{π} have the same law now completes the proof.

We now proceed to the

Proof of Theorem II.1. Start with an arbitrary policy $\pi \in \Gamma_0$ acting on the system in an initial state z. By Lemma II.1, we can construct an alternative policy $\pi_1 \in \Gamma_0$ which schedules according to the STE-rule at the first decision instant along its trajectory and which satisfies the relation $L^{\pi_1} \leq_{st} L^{\pi}$. We proceed inductively; that is, by repeating the same construction n times we can define a policy $\pi_n \in \Gamma_0$ which schedules according to the STE-rule at least at the first n decision points along its trajectory and satisfies

$$L^{\pi_n} \leq_{st} L^{\pi_{n-1}} \leq_{st} \cdots L^{\pi_1} \leq_{st} L^{\pi}.$$

Fix $x \in I\!R$,a positive integer$k$ and $t_i \in [0, \infty), 1 \leq i \leq k$ and pick $g: I\!R^k \mapsto I\!R$. Let A^{γ} denote the event $\{g(L_{t_1}^{\gamma}, \cdots L_{t_k}^{\gamma}) > x\}$ for a policy $\gamma \in \Gamma_0$. Let $t_j = \max_{1 \leq i \leq k} t_i$ and take $\{S_n\}_1^{\infty}$ to be the service times of the messages in the system. Since the policies STE and π_n agree on their first n decisions, one has for all n,

$$\begin{aligned}
P(A^{ste}) &= P(A^{ste} \cap \{\sum_{i=1}^{n} S_i \geq t_j\}) + P(A^{ste} \cap \{\sum_{i=1}^{n} S_i < t_j\}) \\
&\leq P(A^{\pi_n} \cap \{\sum_{i=1}^{n} S_i \geq t_j\}) + P(\sum_{i=1}^{n} S_i < t_j) \\
&\leq P(A^{\pi_n}) + P(\sum_{i=1}^{n} S_i < t_j) \\
&\leq P(A^{\pi}) + P(\sum_{i=1}^{n} S_i < t_j).
\end{aligned}$$

Taking the limit as $n \nearrow \infty$, it now follows that $P(A^{ste}) \leq P(A^{\pi})$, that is, $L^{ste} \leq_{st} L^{\pi}$.

Remarks.

1. The theorem is not true when the assumption of exponential service times is relaxed.

Counterexample:

Consider a single server queueing system in which service times are deterministic and identically equal to 5 units, and there are exactly 5 arrivals into the system at times $3, 4, 12, 16,$ and 16.5 units with corresponding deadlines $3, 9, 2, 9$ and 3.5 units. Let a message with extinction time 3 units and residual service time 5 units be initially present in the system. Then a comparison of the Last-Come-First-Served (LCFS) and STE policies shows

$$E[L_t^{lcfs}] = L_t^{lcfs} = 1(t \geq 6),$$

$$E[L_t^{ste}] = L_t^{ste} = 1(t \geq 14) + 1(t \geq 20),$$

where $1(A)$ is the indicator function of the event A. Thus,

$$E[L_t^{lcfs}] < E[L_t^{ste}] \qquad \forall t \geq 20.$$

2. A close look at the proof also shows that the exponentiality of the service times was used only to match service completions of arrivals to those who have already started service (e.g case 2(b)). Also, idling doesn't pay when there are no arrivals into the system after it has started operation. So, if we consider a single server queue under assumption A1 such that all the messages are initially present in the system and there are no additional arrivals, following the same line of reasoning as in the Theorem, we obtain the (stronger) result

$$L^{ste}(z) \leq_{st} L^{\pi}(z) \qquad \forall \pi \in \Gamma_1, \forall z \in Z.$$

Next, we consider optimality within Γ_1, the broader class of only nonpreemptive policies. Now, the idling of the server is allowed. In this case, examples can be easily constructed to show that the STE policy is no longer optimal; the basic intuition being that when all the messages awaiting service have large extinction times, it pays to idle in expectation of a message with a very short deadline. However, the philosophy of STE-scheduling still plays an important role as the following result demonstrates.

Proposition II.1 *Consider a single server queue under assumption A1. Then, for every policy* $\pi \in \Gamma_1$, *there exists a policy* $\hat{\pi} \in STEI$, *such that*

$$L^{\hat{\pi}}(z) \leq_{st} L^{\pi}(z) \qquad \forall z \in Z.$$

As in the previous theorem, the proof can be worked out in two steps. In the first step, assuming that t_0 is arbitrary decision instant at which π schedules a message that is not the one with the smallest time to extinction, we construct, using the knowledge of π, a policy $\tilde{\pi}$, which schedules the message with the smallest time to extinction and satisfies $L^{\tilde{\pi}} \leq_{st} L^{\pi}$. This then can be used recursively to improve upon any given policy in Γ_1 until the improving policy belongs to the class of $STEI$ policies. The key observation which in fact facilitates the arguments in the first step is that when idling is permitted, the policy $\tilde{\pi}$, which we construct, can be chosen to follow π at all times beyond t_0. Exponentiality of the service times is therefore not needed. We omit the details. Under the same assumptions, [5] contains a different proof for the optimality of $STEI$ policies for long run expected fraction of successful customers.

Consider now a special case of interest that involves the following assumption:

(A2) The deadlines associated with the messages are deterministic and identical for all messages, that is, $d_i = d$ for all i. Also the messages initially present in the system all have extinction times less than or equal to d.

As a simple consequence of the above proposition, one now has the following

Corollary II.1. *Consider a single server queueing system under assumptions A1 and A2. Then for every initial state $z \in Z$, we have,*

$$L^{ste}(z) \leq_{st} L^{\pi}(z) \qquad \forall \pi \in \Gamma_1.$$

Proof. First note that by Proposition II.1, it suffices to prove the claim for $\pi \in$ STEI. We proceed as before. Let t_0 be an instant at which a policy $\pi \in$ STEI chooses to idle. It now suffices to construct $\tilde{\pi} \in \Gamma_1$ which schedules the customer with the shortest time to extinction at time t_0 and satisfies $L^{\tilde{\pi}} \leq_{st} L^{\pi}$.

Let $E(t_0) = \{e_1 \cdots e_n\}$ with $n \geq 1$. Because of assumption A2, it is clear that the arrivals in (t_0, ∞) have extinction times no smaller than e_n. Let τ be the first instant after t_0 at which π decides to schedule a message. Consider the following two cases :

1). Suppose $\tau < e_1$. Since $\pi \in STEI$, π schedules e_1 at time τ. Take the first service time under π (which begins at time τ) to be the same as the service time of e_1 under $\tilde{\pi}$.Following its own trajectory, therefore, $\tilde{\pi}$ can determine the instant of completion of service of e_1 under π. Let $\tilde{\pi}$ idle from the time the service of e_1 is finished under $\tilde{\pi}$ to the time of completion of service of e_1 under π. Letting $\tilde{\pi}$ follow π from this time onwards, one obtains $\bar{L}_t^{\pi} = L_t^{\tilde{\pi}}\ \forall t \geq t_0$.

2). Suppose $\tau \geq e_1$. Let π schedule e^* at time τ. Take the service time of e^* under π to be equal to that of e_1 under $\tilde{\pi}$. Let $\tilde{\pi}$ never schedule e^* and instead idle, if necessary, as in the previous case. It is then clear that

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [t_0, e_1) \cup [e^*, \infty); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_1, e^*). \end{cases}$$

Because of Assumption A1, L^{π} and $\bar{L}^{\pi}$ have the same distribution, and so, $L^{\tilde{\pi}} \leq_{st} L^{\pi}$.

III. CONSTRAINTS ON COMPLETE TRANSMISSION TIMES

We now assume that messages have limitations on their entire sojourn times in the system. Each message upon its arrival into the system at time t_i declares a deadline d_i such that, if by its extinction time $e_i = t_i + d_i$ its transmission is not completed, the message is considered lost. If its extinction time occurs while it is awaiting transmission, it is never scheduled. It may also happen that the message is in the process of being transmitted when its extinction time occurs. We *assume* in this case that its transmission is aborted at that moment and it is considered lost.

We first consider optimality within the class of nonpreemptive nonidling policies. The following parallels Theorem II.1 closely. We provide the proof in its entirety but in a somewhat terser fashion.

Theorem III.1 *Consider a single server queue under assumption A1 and exponential service time distribution. Then for every initial state $z \in Z$, we have,*

$$L^{ste}(z) \leq_{st} L^{\pi}(z) \qquad \forall \pi \in \Gamma_0.$$

Proof. To prove the result, we follow the same program as illustrated in Theorem II.1. With $E(t_0) = \{e_1 \cdots e_n\}$ $(n \geq 2)$, let $\pi \in \Gamma_0$ be a policy which schedules message e_k $(2 \leq k \leq n)$ at time t_0. Using the knowledge of π, we construct $\tilde{\pi}$ which schedules e_{k-1} at time t_0 and satisfies $L^{\tilde{\pi}} \leq_{st} L^{\pi}$. First an induction on k and then an induction on the decision instants will provide the final result.

Consider the systems evolving under policies π and $\tilde{\pi}$. Couple the realizations by giving them the same arrival and deadline processes in $[t_0, \infty)$. Let σ be the instant at which the service which begins at time t_0 under $\tilde{\pi}$ would complete if uninterrupted. Take the service time of e_k under π to be equal to that of e_{k-1} under $\tilde{\pi}$. Two cases exhaust the possibilities:

Case 1 : $\sigma \geq e_{k-1}$.

Consider the interval $[t_0, e_{k-1}]$. All arrivals in this interval having extinction times less than or equal to e_{k-1} and messages $e_1 \cdots e_{k-2}$ (if $k > 2$) are all lost under both policies as none of them could commence transmission. Message e_{k-1} is lost under π as it could not begin transmission and under $\tilde{\pi}$ as it could not complete transmission. At time e_{k-1}, schedule e_k under $\tilde{\pi}$ and assign the residual service time of e_k under π to be equal to the new service time of e_k under $\tilde{\pi}$. Again, this becomes possible because service times are independent and exponentially distributed with the same rate. Also, let $\tilde{\pi}$ follow π hence onwards; clearly this leads to the conclusion that $\bar{L}_t^\pi = L_t^{\tilde{\pi}}$ $\forall t \geq t_0$.

Case 2 : $\sigma < e_{k-1}$.

At time σ, the sets of extinction times under the two policies differ only in that e_{k-1} is included in that set under π as compared to e_k under $\tilde{\pi}$. Let $\tilde{\pi}$ follow π after σ except that it schedules e_k when (and if) π schedules e_{k-1}. Thus $\tilde{\pi}$ is well defined in $[\sigma, \tau)$, where τ is the end of the current busy period under π. We now have three cases to consider depending on whether π scheduled e_{k-1} or not, and whether, when scheduled, the service completed or not. In what follows, we treat only the case in which π never managed to schedule e_{k-1}. Other cases can be handled similarly and we omit their discussion.

Suppose π never scheduled e_{k-1} for service in $[\sigma, e_{k-1})$. Clearly $\tau \geq e_{k-1}$ as π is nonidling. If $\tau \geq e_k$ as well, letting $\tilde{\pi}$ follow π after time τ , it follows that,

$$\bar{L}_t^\pi = L_t^{\tilde{\pi}} + 1\{e_{k-1} \leq t < e_k\} \quad \forall t \geq t_0.$$

Suppose now that $\tau < e_k$. Schedule e_k under $\tilde{\pi}$ at time τ. If there are no arrivals while e_k is in service under $\tilde{\pi}$, we let $\tilde{\pi}$ follow π from time e_k onwards and conclude

$$\bar{L}_t^\pi = L_t^{\tilde{\pi}} + 1\{e_{k-1} \leq t < e_k\} + 1\{t \geq e_k, \sigma_1 < e_k\},$$

where σ_1 is the time at which e_k leaves the system under $\tilde{\pi}$.

Consider now the case in which there is at least one arrival in $[\tau, \sigma_1)$ and the first message that starts service under π has extinction time e_b. Assign the service of e_b under π to be equal to the residual service time of e_k under $\tilde{\pi}$. Let σ_2 be the instant at which the above service would have completed if uninterrupted. Consequently we have to discuss the following cases:

1) Suppose $e_k < e_b$ and $\sigma_2 \geq e_k$. Then upon scheduling e_b under $\tilde{\pi}$ at time e_k and assigning the remaining service time under π to be equal to the new service time of e_b under $\tilde{\pi}$, and letting $\tilde{\pi}$ follow π hence onwards, we obtain,

$$\bar{L}_t^\pi = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [t_0, e_{k-1}) \cup [e_k, \infty), \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, e_k). \end{cases}$$

2) Suppose now that $e_b < e_k$ and $\sigma_2 \geq e_b$. Then, at time e_b, the message e_b is lost under both policies. If no messages begin service under π in $[e_b, \sigma_2)$, letting $\tilde{\pi}$ follow π from time σ_2 onwards, we have,

$$\bar{L}_t^{\pi} = L_t^{\tilde{\pi}} + 1\{e_{k-1} \leq t < e_k\} + 1\{t \geq e_k, \sigma_2 < e_k\}.$$

On the other hand, suppose a message begins service under π in $[e_b, \sigma_2)$. This is a situation we have already discussed. By repeating the arguments in an obvious way, we are led to $\bar{L}_t^{\pi} \geq L_t^{\tilde{\pi}}\ \forall t \geq t_0$.

3) Consider the case $\sigma_2 < \min(e_b, e_k)$. Let $\tilde{\pi}$ follow π for $t \geq \sigma_2$ and assume again that the corresponding services are of equal duration. Thus $\tilde{\pi}$ is well defined in $[\sigma_2, \tau_1)$, where now τ_1 is the end of this current busy period under π. If $\tau_1 \geq e_b$, one has

$$\bar{L}_t^{\pi} = \begin{cases} L_t^{\tilde{\pi}} & \text{if } t \in [t_0, e_{k-1}) \cup [e_b, \infty); \\ L_t^{\tilde{\pi}} + 1 & \text{if } t \in [e_{k-1}, e_b). \end{cases}$$

For $\tau_1 < e_b$, schedule e_b under $\tilde{\pi}$ and proceed as discussed before to obtain $\bar{L}_t^{\pi} \geq L_t^{\tilde{\pi}}\ \forall t \geq t_0$.

4) If $e_b = e_k$ and $\sigma_2 \geq e_b = e_k$, letting $\tilde{\pi}$ follow π from time $e_b = e_k$ onwards, one has $\bar{L}_t^{\pi} = L_t^{\tilde{\pi}}\ \forall t \geq t_0$.

Remarks.

1. The theorem is not true without the exponential service assumption.

Counterexample: Consider a single server queueing system in which two messages with extinction times 2 and 6 units are initially present in the system. The service times form a renewal process with the common distribution : $P(S_1 = 0) = p, P(S_1 = 5) = q = 1-p$. Assume also that $0 < p < q$. Let LTE be the policy which serves the message with the largest time to extinction. Elementary calculations show that for all $t \geq 6$,

$$\bar{L}_t^{lte} - L_t^{ste} = \begin{cases} 1 & \text{with } \textit{prob. } pq; \\ 0 & \text{with } \textit{prob. } p^2 + pq; \\ -1 & \text{with } \textit{prob. } q^2. \end{cases}$$

where $\bar{L}^{lte}$ is obtained by coupling the service times to those for L^{ste}. Since $\bar{L}^{lte}$ has the same law as L^{lte}, it follows that for all $t \geq 6$,

$$E[L_t^{lte}] - E[L_t^{ste}] = pq - q^2 = (p - q)q < 0.$$

2. Consider now the situation, as in Remark 2 following Theorem II.1, of a single server queue under assumption A1 when arrivals into the system are disallowed after the system has started operation. Following the proof of Theorem III.1, it is clear that under the *additional* assumption of exponential service times, STE is optimal in the class Γ_1 of nonpreemptive policies. The example in the previous remark shows the necessity of this additional assumption. This is in contrast to the result in section II.

We now consider optimality within Γ_1, the broader class of nonpreemptive policies that permit idling. A careful examination of the previous example shows that with only assumption A1, the optimal policy need not belong to the class of STEI policies. Also, a slight modification of the above example shows that the STE policy need not be optimal under assumptions A1 and A2. These observations should be compared to the conclusions of Proposition II.1 and its Corollary. However, we can prove the following:

Proposition III.1 *For a single server queue under assumptions A1 and exponential service time distributions, for every $\pi \in \Gamma_1$, there exists $\tilde{\pi} \in STEI$ such that for each initial state $z \in Z$, $L^{\tilde{\pi}}(z) \leq_{st} L^{\pi}(z)$.*

Corollary III.1 *For a single server queue as above and the added assumption A2, we have, $L^{ste}(z) \leq_{st} L^{\pi}(z)$ $\forall \pi \in \Gamma_1$, $\forall z \in Z$.*

The proofs are almost verbatim versions of those of Proposition II.1 and its Corollary with minor modifications and hence are omitted.

Let us now consider policies that are allowed to preempt messages if necessary, and consider optimality within Γ, the class of arbitrary (nonanticipative) policies. Let STE(P) denote a preemptive version of the STE policy, that is, a policy in Γ which *always* transmits the message with the shortest time to extinction; it interrupts current transmission when there is an arrival with an earlier extinction time and commences transmission of this message.

We have the following result :

Theorem III.2 *Consider a single server queue under assumption A1. Assume also that the common distribution of the service times is exponential. Then, for each initial state $z \in Z$, we have $L^{ste(p)}(z) \leq_{st} L^{\pi}(z)$ $\forall \pi \in \Gamma$.*

Proof. To avoid repetitiveness, we will only outline the proof. Consider any policy $\pi \in \Gamma$. Let $E(t_0) = \{e_1 \cdots e_n\}$ where t_0 is an arbitrary decision instant. Assume that at t_0, either $n \geq 2$ and π schedules e_k $(2 \leq k \leq n)$ or $n \geq 1$ and π chooses to idle . We will construct $\tilde{\pi}$ which schedules e_1 at time t_0, takes the next action only at an event epoch (i.e. at an instant of an arrival, service completion or the loss of a message) and satisfies $L^{\tilde{\pi}} \leq_{st} L^{\pi}$. Usual induction arguments then complete the proof.

Consider the parallel evolution of the system under the two policies $\bar{\pi}$ and π and couple the realizations by giving them the same arrival and deadline processes.

Consider the case when $n \geq 2$ and π schedules e_k $(2 \leq k \leq n)$.Depending on the nature of the next event, the following cases have to be discussed:

(a) The next event is a service completion (under both π and $\bar{\pi}$). Let $\bar{\pi}$ follow π henceforth, except that (i) it schedules e_k when (and if) π schedules e_1, and (ii) it either never schedules or preempts and never reschedules e_k if e_1 is lost under π. The usual arguments then lead to the relation $\bar{L}_t^{\bar{\pi}} \leq L_t^{\pi}$ $\forall t \geq t_0$.

(b) The next event is either the loss or the arrival of a message (under π and $\bar{\pi}$). Letting $\bar{\pi}$ to be identical to π hence onwards, we have $\bar{L}_t^{\pi} = L_t^{\bar{\pi}}$ $\forall t \geq t_0$.

(c) Suppose that π simply preempts e_k without any other event happening. If π schedules e_1, define $\bar{\pi}$ from this instant onwards to follow π. Otherwise, we are in a situation similar to that at time t_0, and the arguments can be repeated.

Consider now the case when $n \geq 1$ and π decides to idle. The important point here is that the because of the coupling, the first time π decides to schedule a message can be determined following the realizations of $\bar{\pi}$. The proper actions of $\bar{\pi}$ can then be stipulated to lead to the desired result. The accounting is straight forward but very repetitive and hence is omitted.

The proof is then completed by noting that $\bar{L}_t^{\pi}$ was chosen to have the same distribution as L_t^{π}.

Remarks.

1. The counterexample following Theorem III.1 demonstrates the necessity of the assumption of exponential service time distribution.

2. From the proof it is clear that idling never pays under assumption A1 when premption is allowed.

3. Let $T^{\pi}(z) = \{T_k^{\pi}(z), k = 1, 2, \cdots\}$ where $T_k^{\pi}(z) \stackrel{\text{def}}{=} \inf\{t \geq 0 : L_t^{\pi}(z) \geq k\}$. for $k = 1, 2, \cdots$ be the discrete time process representing the instants at which messages are lost under a policy π when starting from state z at time 0. Since

$$T^{\pi_1}(z) \geq_{st} T^{\pi_2}(z) \iff L^{\pi_1}(z) \leq_{st} L^{\pi_2}(z),$$

for any two policies π_1 and π_2, all results in Sections II and III are valid for the cost T_k^{π} with the obvious modifications. In reliability applications [2], the system is often considered to have failed when the first deadline is missed. The problem of stochastically maximizing the RV $T_1^{\pi}(z)$ is then appropriate. It is interesting to observe that for the optimality of STE in this case, exponentiality of service times is not needed in Theorem II.1 and in all the results of this section.

Weaker versions of some of the results above were presented in [4]. In that paper, optimality with respect to the long run expected number of successful customers was considered. A theorem analogous to our Theorem III.1 was proven when arrivals are Poisson and optimality is considered within the class of stationary policies. Statements analogous to our Proposition III.1 and Theorem III.2 are also given under an additional hypothesis (Assumption 4.2 in [4], Page 71), which seems difficult to verify. Furthermore, the arguments and methodology used there are substantially different from ours.

IV. SCHEDULING UNDER A REDUCED INFORMATION STRUCTURE

In this part, we re-examine the problem considered in the last two sections. We assume that due to implementational difficulties, the scheduler does not have precise knowledge of the extinction times of the messages. The information regarding the distribution of the deadlines is however available and there is a mechanism which informs the scheduler when messages are lost. Furthermore, the extinction times become available as soon as a message begins service and actions can be taken as in the previous sections. Interestingly enough, providing the scheduler with less detailed information tends to simplify the situation. Under some reasonable assumptions on the deadlines, idling policies and the exponentiality of service times play a less important role in the analysis and reasonably complete results are obtained.

Consider the following single server queueing system. Customers arrive at times $\{t_i\}$ and depart at times $\{t_i + d_i\}$ if service doesn't begin (or isn't completed). Only the common distribution $F(.)$ of $\{d_i\}$ is available to the scheduler, but the scheduler is also *aware* of the departures of customers when they occur. Assume that $F(.)$ has a density $f(.)$; it is well known that the *failure rate function* $\lambda(t) \overset{\text{def}}{=} f(t)/(1-F(t))$ (defined for all those values of t for which $F(t) < 1$) has a useful probabilistic interpretation; namely $\lambda(t)dt$ is the probability that the RV having d.f $F(.)$ takes value in $[t, t+dt]$ given that it is greater than t. It seems natural to assume that beyond a certain time, the impatience of the customers doesn't decrease. We, therefore, make the following assumption :

(A3) $\{d_i\}_{i=1}^{\infty}$ is a sequence of i.i.d RV's which are independent of $\{T_i\}_{i=1}^{\infty}$. Furthermore the common distribution has a non-decreasing failure rate function.

We first consider the situation when deadlines are to the beginning of service. A customer is termed eligible at time t if it has arrived but neither begun service nor reneged by t. Let $W(t)$

denote the (decreasing) ordered set of waiting times of the eligible customers at time t. In Renewal Theoretic terminology, $W(t)$ is the ordered set of their ages.With $H_a(t)$, $C_s(t)$, and $i(t)$ as defined in Section II, $z(t) = (W(t), H_a(t), C_s(t), i(t))$ is a suitable state description under assumptions $A1$ and $A3$. Let Z be its allowable range of values.

Let EA denote the policy in Γ_0 which schedules, at every decision instant, the eligible customer with the earliest arrival instant. This differs from the well known FCFS policy because of possible non-zero initial conditions. Let EAI denote the class of policies in Γ_1 which are allowed to idle when customers are waiting, but which schedule according to EA mechanism when they choose not to idle.

We approach our main results (Theorems IV.1 and IV.2) through the following results. The first lemma indicates the way in which the non-decreasing failure rate assumption on the deadline distribution is going to be used.

To state the result, we introduce some notation. For a nonnegative RV X, let X_t denote its residual life after t units, that is,

$$P(X_t > z) = P(X > z + t \mid X > t).$$

Let $\{X^{(i)}\}_{i=1}^{n}$ be a sequence of i.i.d nonnegative RV's. It is easy to show that $X_{t_1}^{(1)}, X_{t_2}^{(2)}, \cdots X_{t_n}^{(n)}$ are independent (and identically distributed if $t_i = t \; \forall i$). If $\{X^{(i)}\}_{i=1}^{n}$ is independent of a σ-algebra of events $\mathcal{F}$, then so is $\{X_{t_i}^{(i)}\}_{i=1}^{n}$. Furthermore,

Lemma IV.1 [14, Prop. 8.1.3] *If the common distribution has nondecreasing failure rate, then for every $t_1 \geq t_2 \geq \cdots \geq t_n$,*

$$X_{t_1}^{(1)} \leq_{st} X_{t_2}^{(2)} \leq_{st} \cdots \leq_{st} X_{t_n}^{(n)}.$$

The next lemma, in the same spirit as Propositions II.1 and III.1, characterizes a good set of scheduling policies within the class of admissible ones.

Lemma IV.2 *Consider a single server queue under assumptions A1 and A3. Then for every policy $\pi \in \Gamma_1$, there exists a policy $\hat{\pi} \in EAI$, such that*

$$L^{\hat{\pi}}(z) \leq_{st} L^{\pi}(z) \;\; \forall z \in Z.$$

Proof. Since the arguments closely parallel those in the Section II, we will only point out the essential differences. Let $W(t_0) = \{w_1, w_2, \cdots w_n\}$ $(n \geq 2)$ and suppose that π schedules w_k $(2 \leq k \leq n)$ at an arbitrary decision instant t_0. We first construct $\tilde{\pi} \in \Gamma_1$ which schedules w_{k-1} at time t_0 and satisfies $L_{\tilde{\pi}} \leq_{st} L_{\pi}$. Couple the realizations under π and $\tilde{\pi}$ so that

a) the arrivals are the same,

b) $\bar{e}_i^{\pi} = e_i^{\tilde{\pi}}$ $\forall i \neq k, k-1$, $\bar{e}_{k-1}^{\pi} \leq e_k^{\tilde{\pi}}$ *a.s.* and $\bar{e}_{k-1}^{\pi} =_{st} e_{k-1}^{\pi}$, where e_i^{π} is the extinction time of w_i under policy π,

and c) the deadlines (and hence the extinction times) are the same for all arrivals in (t_0, ∞).

Note that the construction as stated in b) is possible because of Lemma IV.1 and the observations stated before the lemma. The coupling for service times and the suitable definition for $\tilde{\pi}$ in (t_0, ∞) can now be given using the ideas in Section II and the rest of the argument also follows easily.

We show next that idling doesn't help. This will then further reduce the good set of scheduling strategies to the EA strategy. Towards this end, start with a policy $\pi \in EAI$ and let t_0 be an instant at which π idles. With $W(t_0) = \{w_1, \cdots w_n\}$ $(n \geq 1)$, let $\tilde{\pi}$ be another policy which schedules w_1 at time t_0. Let τ be the the first instant after t_0 at which π schedules a customer and σ be the time at which w_1 completes service under $\tilde{\pi}$. Couple the realizations under π and $\tilde{\pi}$ to have the same arrival and deadline processes. Note that τ can then be determined observing the realization of $\tilde{\pi}$. Now, if $\tau \geq \sigma$, letting $\tilde{\pi}$ follow π from σ onwards (by inserting idle periods if necessary), one concludes, $\forall t \geq t_0$, that

$$\bar{L}_t^{\pi} = L_t^{\tilde{\pi}} + 1(t \geq e_1, e_1 \leq \tau).$$

When $\tau < \sigma$, the argument proceeds as follows. Let π begin serving w^* (with extinction time e^*) at τ. Take the first service time under π and $\tilde{\pi}$ to be equal. Also, let $\tilde{\pi}$ idle in $[\sigma, \sigma + \tau]$, and follow π from $(\sigma + \tau)$ onwards, except that it never serves w^* in $[\sigma + \tau, \infty)$. Recalling that $\pi \in EAI$, it follows that, $\forall t \geq t_0$,

$$\bar{L}_t^{\pi} = L_t^{\tilde{\pi}} + 1(e_1 \leq t < e^*, w_1 \neq w^*).$$

We have thus shown the following

Theorem IV.1 *For a single server queue under assumptions A1 and A3,*

$$L^{ea}(z) \leq_{st} L^{\pi}(z) \qquad \forall \pi \in \Gamma_1, \forall z \in Z.$$

When deadlines are to the completion of service, the following can be easily shown combining the ideas given so far.

Theorem IV.2 *Consider a single server queue under assumptions A1 and A3. Assume further that the common distribution of the service times is exponential. Then*

$$L^{ea}(z) \leq_{st} L^{\pi}(z) \qquad \forall \pi \in \Gamma_1, \forall z \in Z.$$

Finally, we observe that in sections II, III and IV, one could also have considered the problem of maximizing the number of successful customers. With obvious modifications, the optimality results extend to this situation as well.

V. RESULTS FOR A TANDEM NETWORK

In this section we consider a network of M communication links in tandem. Messages arrive at times $\{t_i\}_{i=1}^{\infty}$ at the first link with corresponding deadlines $\{d_i\}_{i=1}^{\infty}$; thus they carry extinction times $\{e_i = t_i + d_i\}_{i=1}^{\infty}$ which become known to the scheduler upon arrival. We make the crucial assumption that each message must be transmitted in all of the M tandem links, even if its extinction time expires while in transit or prior to transmission. Therefore messages are *not* lost and it seems more appropriate to refer to extinction times as due dates as we will do in the rest of this section.

Let the various classes of scheduling policies be as defined in section II. The control action involves deciding at appropriate decision instants at each of the links which message to transmit, if at all, out of the pool of currently available messages at that link.

Consider operating the above system over the time interval $[0, T]$. Let C_i^{π} be the time at which the message arriving at time t_i with deadline d_i completes service, under policy π, at the last of the M links and departs from the network. Set $C_i^{\pi} = T$ if the message does not depart by time T. Let $h : I\!R \mapsto I\!R$ be a continuous convex function with $h(x) = 0 \;\; \forall x \leq 0$. When starting from state z, the total average cost incurred by operating the system $[0, T]$ under policy π is thus $E[\sum_{i=1}^{N} h(C_i^{\pi}(z) - e_i)]$ where $N = N(T) \overset{\text{def}}{=} \max\{i : t_i \leq T\}$ is the number of arrivals by time T. We are interested in characterizing the scheduling policy which minimizes the above cost.

We make the following assumption :

(A4) The service times at each node form a sequence of i.i.d RV's that are independent of each other, of the arrivals at the first node, and of the deadlines. Also, the arrival process to the first node is nonexplosive.

We only consider optimality within Γ_0, the class of nonpreemptive and nonidling policies. Let the Earliest Due Date (EDD) policy denote the policy in Γ_0, which at a decision instant at any node in the network schedules the message with the smallest due time.

Our main result is the following

Theorem V.1. *Consider the tandem network as described above under assumption A4. Let $h : I\!R \mapsto I\!R$ be a continuous convex mapping with $h(x) = 0 \ \forall x \leq 0$. Then*

$$E\left[\sum_{i=1}^{N} h\left(C_i^{edt}(z) - e_i\right)\right] \leq E\left[\sum_{i=1}^{N} h\left(C_i^{\pi}(z) - e_i\right)\right] \qquad \forall \pi \in \Gamma_0.$$

Proof. The proof again uses essentially the same ideas as in Sections II and III. Let $E(t_0) = \{e_1 \cdots e_n\}$ be the set of due times of the messages awaiting service at node I ($1 \leq I \leq M$) , and suppose that π schedules e_k ($2 \leq k \leq n$) at time t_0, an arbitrary decision instant at that node. We will construct $\tilde{\pi}$ which schedules e_{k-1} at time t_0 and satisfies $J^{\tilde{\pi}}(z) \leq J^{\pi}(z)$,where $J^{\pi}(z) = E[\sum_{i=1}^{N} h(C_i^{\pi}(z) - e_i)]$. Usual induction arguments will then complete the proof. We drop z for notational convenience.

Couple the realizations of the system evolving under π and $\tilde{\pi}$ by giving them the same arrival and deadline processes. Take the service time of e_k at node I under π to be the same as that of e_{k-1} at node I under $\tilde{\pi}$.

Consider first the case $I < M$. Let $\tilde{\pi}$ follow hence onward all actions of π at nodes $1, 2, \cdots, I$ except that at node I, $\tilde{\pi}$ serves e_{k-1} while π serves e_k. Consider the actions of $\tilde{\pi}$ at the subsequent nodes $I+1, \cdots, M$. Let $\tilde{\pi}$ follow all actions of π regarding messages other than e_k and e_{k-1}. About actions regarding e_k and e_{k-1}, consider the following two subcases:

1. Suppose that at the J^{th} node $(I + 1 \leq J \leq M)$, π prefers to schedule e_{k-1} when e_k is waiting at the same node. Let $\tilde{\pi}$ then follow π regarding messages e_k and e_{k-1} as well; this will result in equal costs incurred under the two policies.

2. Suppose, on the other hand, that π prefers to schedule e_k before e_{k-1} at each of the nodes $I + 1 \leq J \leq M$. Let $\tilde{\pi}$ serve e_{k-1} when π serves e_k and viceversa. First note that $\bar{C}_i^{\pi} = C_i^{\tilde{\pi}} \ \forall i \neq k, k-1$ and $U \stackrel{\text{def}}{=} \bar{C}_k^{\pi} = C_{k-1}^{\tilde{\pi}} \leq V \stackrel{\text{def}}{=} \bar{C}_{k-1}^{\pi} = C_k^{\tilde{\pi}}$ for all sample paths and hence

$$\begin{aligned}\Delta &= \sum_{i=1}^{N}[h(\bar{C}_i^{\pi} - e_i) - h(C_i^{\tilde{\pi}} - e_i)] \\ &= h(V - e_{k-1}) + h(U - e_k) - h(U - e_{k-1}) - h(V - e_k).\end{aligned}$$

The assumed properties of h imply that h is nonnegative and nondecreasing and that

$$h(x) + h(y) \geq h(z_1) + h(z_2),$$

for any z_1, z_2 $\in [x, y]$ such that $\min(z_1, z_2) - x \leq y - \max(z_1, z_2)$. This immediately shows that $\Delta \geq 0$.

When $I = M$, the policy $\tilde{\pi}$ is defined in a similar fashion; we omit the details since the discussion will be verbatim. Since C_i^{π} has the same distribution as $\tilde{C}_i^{\pi}$ for each i, the conclusion follows.

Remarks.

The optimal scheduling policy just obtained has a nice distributed implementation; actions which are globally optimal can be taken at each queue based just on the due times of the messages waiting at that queue only.

VI. CONCLUSIONS

We have considered the problem of dynamically scheduling the transmission of messages in situations where messages have individual constraints on either their waiting time or complete transmission times. First, a single node was considered in which the messages were lost if the constraints were not met. Different classes of admissible policies were considered. Under nonexplosive, but otherwise arbitrary, arrival and completely arbitrary deadline processes, but under some restrictions on service times, the intuitively appealing policy of scheduling the "most urgent" message was shown, in many cases, to minimize the number of messages lost in a strong pathwise sense. A tandem network of M nodes was then considered. We assumed that messages are never lost, but a penalty is incurred if a message doesn't complete service in the tandem by its extinction time (also called due time in this case). Under fairly general restrictions on arrival, deadline and service processes, the intuitive policy of scheduling, at each node, the message which is "earliest due" is shown to minimize a tardiness cost over a finite operating horizon among all nonidling, nonpreemptive policies. Simple interchange arguments were used to establish all our results. As is clear in the development of the paper, the particular application on which we concentrated is only incidental in that the entire approach (model formulation and solution) is quite general and applies to abstract queueing systems with customers and services of unspecified application features.

VII. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Pantelis Tsoucas for helpful discussions.

VIII. REFERENCES

[1] F. Baccelli, P. Boyer, G. Hebuterne, "Single Server Queues with Impatient Customers," *Adv. Appl. Prob.*, Vol. 16, pp. 887-905 (1984).

[2] F. Baccelli, K.S. Trivedi, "A Single Server Queue in a Hard Real Time Enviornment," *Oper. Res. Letters* , Vol. 4, No.4, pp. 161-168 (1985).

[4] B.T. Doshi, " An M/G/1 Queue with a Hybrid Discipline," *The Bell Syst. Tech. Jour.*, Vol. 62, No. 5, pp. 1251-1271 (1983)

[4] S.S. Panwar, "Time-constrained and Multiaccess Communications," Ph.D. dissertation, University of Massachusetts, Amherst (1985).

[5] S.S. Panwar, D. Towsley, J.K. Wolff, "Optimal Scheduling Policies for a Class of Queues with Customer Deadlines to the beginning of Service," *Jour. Assoc. Comp. Mach.*, Vol. 35, No. 4, pp. 832-844 (1988).

[6] P.R. de Waal, "Performance Analysis and Optimal Control of an $M/M/1/K$ Queueing System with Impatient Customers," Research Report $OS-R8713$, Stichting Mathematisch Centrum, Amsterdam (1987).

[7] M. Pinedo, "Stochastic Scheduling with Release Dates and Due Dates," *Oper. Res.* , Vol. 31, pp. 559-572 (1983).

[8] Z.S. Su, K.C. Sevick, "A Combinatorial Approach to Scheduling Problems," *Oper. Res.* , Vol. 26, pp. 836-844 (1978).

[9] J.R. Jackson, "Scheduling a Production Line to Minimize Maximum Tardiness," Management Science Research Report 43, UCLA (1955).

[10] J. Walrand, "A Note on Optimal Control of a Queueing System with Two Heterogeneous Servers," *Sys. and Contr. Letters*, Vol. 4, pp. 131-134 (1984).

[11] J. Walrand, *An Introduction to Queueing Networks* , Prentice Hall (1987).

[12] D. Stoyan, *Comparison methods for Queues and other Stochastic Models*, John Wiley & Sons (1983).

[13] T. Kamae, V. Krengel, G.L. O'Brien, "Stochastic Inequalities on Partially Ordered Spaces," *Annals of Prob.*, Vol. 6, No. 6, pp. 1044-1049 (1978).

[14] S. Ross, *Stochastic Processes*, John Wiley & Sons (1983).

[15] P.P. Bhattacharya, A. Ephremides, "Optimal Scheduling of the Transmission of Messages with Strict Deadlines," *Proc. of 1988 Conference on Information Sciences and Systems*, Vol. II, pp. 623-628, Princeton (1988).

[16] Z.J. Wu, P.B. Luh, S.C. Chang, D.A. Castanon, "Optimal Control of a Queueing System with two Interacting Service Stations and Three Classes of Impatient Tasks," *IEEE Trans. on AC*, Vol. 33, No. 1, pp. 42-49 (1988).